← Go to **AAAI 2026 Conference** homepage (/group?id=AAAI.org/2026/Conference)

Sentimentogram: A Personalized Subtitle System Using Multi-modal Emotion Recognition and Visualization

(/pdf?id=K0rDZXTEeS)

Seungkyu Oh (/profile?id=~Seungkyu_Oh2), Kim Gwangsoo (/profile?id=~Kim_Gwangsoo1), Wookey Lee (/profile?id=~Wookey_Lee1)

25 Jul 2025 (modified: 16 Sept 2025) Submitted to AAAI-26 Conference, Area Chairs, Senior Program Committee, Program Committee, Authors Revisions (/revisions?id=K0rDZXTEeS) BibTeX

Serve As Reviewer: Seungkyu Oh (/profile?id=~Seungkyu_Oh2), Wookey Lee (/profile?id=~Wookey_Lee1)
Keywords: Speech Emotion Recognition, Emotion Visualization, Multimodal, Personalized Subtitles
Primary Keyword: APP: Humanities & Computational Social Science
Secondary Keywords: NLP: Sentiment Analysis, Stylistic Analysis, and Argument Mining

Abstract:
Conventional static subtitle systems hinder media immersion by failing to convey emotional nuances and lacking user personalization. This study introduces 'Sentimentogram,' a novel system designed to address these limitations by providing dynamic visualization and personalization of subtitles based on recognized speech emotions.

The core innovation of the system is its emotion visualization framework. It translates discrete recognized emotions into the VAD (Valence-Arousal-Dominance) dimensional space. These VAD coordinates are then systematically mapped to visual properties of the subtitles, such as color hue, saturation, and font weight, enabling an intuitive and nuanced representation of the speaker's affective state. Furthermore, the system delivers a personalized experience by adapting these visual styles to user-specific attributes, such as cultural background.

To power these features, Sentimentogram utilizes a robust multi-modal Speech Emotion Recognition (SER) engine. This engine achieves high accuracy by fusing acoustic features, extracted via a Wav2Vec2 model, with textual features from a BERT model applied to the transcribed speech.

By shifting the focus from mere transcription to a rich, adaptive visual experience, this research demonstrates a significant step toward more immersive and user-centric media consumption. The proposed framework offers a new paradigm for emotionally aware and personalized subtitles.

Country Of Institutions: Korea, Republic of
Supplementary Material: pdf (/attachment?id=K0rDZXTEeS&name=supplementary_material)
Profile Policy Agreement: I confirm that all authors have up-to-date OpenReview profiles, including their current position, institution-affliated email address, and DBLP URL. I understand that submissions with incomplete author profiles will be subject to desk rejection.
Submission Number: 22119

Filter by reply type... Filter by author... Search keywords... Sort: Newest First
Everyone Program Chairs Submission22119... Submission22119... Submission22119... Submission22119... Submission22119... Submission22119... Submission22119... Submission22119... Submission22119...
5 / 5 replies shown

Add: **Withdrawal** **Ethics Chair Author Comment**

Paper Decision

Decision by Program Chairs 15 Sept 2025, 22:01 (modified: 16 Sept 2025, 01:00)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors Revisions (/revisions?id=RXKtRKUDGQ)

Decision: Reject
Comment:
This paper did not advance to Phase 2 review for the AAAI-26 conference.
The paper was reviewed by at least two human reviewers in Phase 1, and unfortunately it did not garner sufficient support for promotion to Phase 2. This decision is final. The provided reviews contain more detail, which we hope will be beneficial to the progress of this research.

Add: **Ethics Chair Author Comment**

Official Review

Official Review by Program Committee NHsn 03 Sept 2025, 16:52 (modified: 22 Sept 2025, 22:53)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee NHsn, Authors Revisions (/revisions?id=chYqESIexK)

Review:
Paper Summary

The paper presents Sentimentogram, a personalized, emotion-aware subtitle generation system. It combines multimodal Speech Emotion Recognition (SER) using Wav2Vec2 and BERT with self and cross-attention fusion. Emotions are mapped to Valence-Arousal-Dominance (VAD) values and visualized in subtitles via color, font, and style changes. The system adds personalization by adapting to age, cultural background, and playback conditions. Experiments show improvements over state-of-the-art SER models on IEMOCAP (76% WA/WF1) and a user study demonstrates gains in emotional reflection, clarity, and satisfaction.

> Strengths of the Paper:

1. The motivation of the paper is clear and valid: conventional subtitles lack emotional nuance.
2. The proposed architecture is technically solid; furthermore, the ablation studies show benefits of attention-based fusion.
3. I like the idea of integrating emotion-color mapping with cultural adaptation.

> Weakness of the Paper:

1. The proposed SER fusion architecture is incremental, cross-attention multimodal fusion is already well-studied.
2. The paper repeatedly claims to be the "first comprehensive effort" at emotion-aware subtitle generation. However, the evaluation relies only on existing SER datasets (IEMOCAP, Condensed Movies) and a small user study. If the work truly represents a first-of-its-kind effort, one would expect either a new dataset or a dedicated evaluation protocol for subtitle personalization. As it stands, the contribution is more a repurposing of SER benchmarks than a new research paradigm.
3. The user evaluation is promising but details are sparse. The number of participants, demographics, and statistical significance of improvements are missing. Without this, the results are hard to interpret.
4. The experimental evaluations are limited, to support author claims broader validations for generalizability are required.

Rating: 4: Ok but not good enough - rejection
Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Add: **Ethics Chair Author Comment**

Sentimentogram Review: Promising Concept for Emotionally Adaptive Subtitles, but Limited Rigor, Real-World Validation, and Accessibility Evidence Suggest Suitability for Demo/Workshop Rather than Top-Tier Venue.

Official Review by Program Committee p457 02 Sept 2025, 04:22 (modified: 22 Sept 2025, 22:53)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee p457, Authors Revisions (/revisions?id=j9wZUe6Yky)

Review:

Review of “Sentimentogram: A Personalized Subtitle System Using Multi-modal Emotion Recognition and Visualization”

Summary

This paper introduces *Sentimentogram*, an innovative system that generates emotion-aware subtitles by fusing acoustic and linguistic features through a multi-modal Speech Emotion Recognition (SER) model. The system dynamically personalizes subtitle visualization—including color, font size, and style—based on detected emotions projected into the Valence-Arousal-Dominance (VAD) space, along with user characteristics such as age, culture, and playback speed. The model achieves competitive performance on the IEMOCAP benchmark and demonstrates improvements in user experience through a preliminary user study.

Strengths:

- **Clear and appealing use case:** Emotion-aware subtitles offer a meaningful enhancement over static subtitles, with potential accessibility and immersion benefits.
- **Sound multi-modal SER architecture:** Combines Wav2Vec2 acoustic features and BERT textual embeddings with self- and cross-attention for effective fusion.
- **Principled VAD-based visualization:** Maps emotions to subtitle visual styles (color, font size, weight), supported by cross-cultural psychology literature.
- **Inclusion of ablation and user evaluations:** Demonstrates performance gains from modality fusion and shows the system’s impact on user engagement and emotional comprehension.

Major Concerns:

Evaluation Limitations:

- **Narrow dataset scope:** Evaluation relies solely on the IEMOCAP dataset (English, acted, small scale), limiting real-world generalizability for multilingual or naturalistic subtitles. Important benchmarks such as CH-SIMS, MOSEI, M3ED, or CREMA-D are missing.
- **No real-world deployment tests:** Lacking assessments of latency, diarization, overlapping speech, and acoustic challenges like background music or environmental noise.
- **Incomplete metrics and protocol transparency:** Reports weighted accuracy (WA) and weighted F1 (WF1) without confidence intervals, statistical significance, or class-balanced metrics like UAR; unclear whether train/test splits are speaker-disjoint; mixing SOTA results with different setups complicates fair comparison.
- **User study under-specified:** No details on sample size, recruitment, demographics, task design, or statistical analysis. Accessibility-relevant groups (hearing-impaired, color-blind) are not considered.

Accessibility & HCI Claims:

- **Potential WCAG violations:** Subtitle color and style adaptations may reduce legibility for color-blind users or create distracting/cognitively taxing displays.
- **Ethical concerns with personalization:** Emoji use and slang substitution risk altering intended meanings; censorship raises consent and ethical questions. Cultural personalization is rule-based without user overrides, risking stereotyping.

Methodological & Design Issues:

- **Labeling rationale unclear:** Starts from discrete Ekman 6+neutral emotion categories but maps to VAD; lacks justification for this indirect approach and no resolution of conflicts between categorical and dimensional representations.
- **Dependence on Whisper STT without robustness analysis:** Potential subtitle errors from ASR mistakes are discussed but not empirically evaluated, which can impact both emotion recognition and user experience.
- **Lack of error analyses:** Missing per-emotion confusion matrices, calibration curves, or assessments of error costs.

Deployment & Practicality:

- **No system latency or computational profile:** No evaluation of FPS, inference time, hardware requirements, or on-device versus cloud processing.
- **Subtitle timing adjustments asserted but unevaluated:** CPS-based splitting and pacing are not benchmarked against standard subtitling guidelines.

Minor Issues:

- Placeholder or outdated references; some figures/tables lack reproducible experiment details (training seeds, splits).
- Absence of ethics review board (IRB) approvals or data privacy discussions for user personalization features.

Recommendations:

- **Broaden evaluation:** Include multiple real-world and multilingual datasets; test streaming scenarios with noise, multi-speaker diarization, and latency/QoS metrics.
- **Robust HCI/accessibility studies:** Recruit diverse user groups (N ≥ 24), including users with hearing/color-vision impairments; apply WCAG guidelines; perform preregistered hypotheses and thorough statistical analyses.
- **Transparent, fair evaluation protocols:** Provide speaker-disjoint splits, confidence intervals, calibration data, and UAR/macro-F1 metrics; analyze the impact of ASR errors systematically.
- **Refined personalization:** Incorporate user-controlled theme customization, color-blind friendly palettes, explainability, and opt-out options.
- **Direct VAD modeling:** Either justify the Ekman → VAD mapping theoretically and empirically or train models to predict VAD dimensions directly for smoother style mapping.
- **Comprehensive deployment reporting:** Include system latency, computational resource use, privacy considerations, and subtitle timing validations.

Summary of the Review:

The *proposed* system demonstrates a promising direction for emotionally immersive, personalized subtitles combining multi-modal SER and contextual adaptation. However, as presented, the paper has **weak to moderate rigor** in evaluation, limited real-world applicability, and underdeveloped HCI accessibility evidence. It needs significant improvements in evaluation breadth, user studies, and transparency before it can be considered for acceptance at a top-tier venue. A demo or workshop presentation after addressing these concerns could be an appropriate next step.

Rating: 4: Ok but not good enough - rejection

Confidence: 5: The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature

Add: **Ethics Chair Author Comment**

An interesting work, but the topic is better suited for Human-Computer Interaction (HCI) journals or conferences

Official Review by Program Committee NUhp 30 Aug 2025, 21:01 (modified: 22 Sept 2025, 22:53)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee NUhp, Authors Revisions (/revisions?id=tdZmacePrx)

Review:

Paper Overview

This paper proposes **Sentimentogram**, a personalized subtitle system based on multimodal emotion recognition.
Key features include:

- Acoustic features extracted using **Wav2Vec2**, linguistic features extracted using **BERT**.
- Multimodal fusion via **self-attention** and **cross-attention**, mapping emotions to the **VAD** space.
- Subtitle appearance (color, size, style) dynamically changes with detected emotion.
- Personalization modules adapt outputs according to user age, cultural background, and playback conditions.
- Experiments show competitive or superior performance on **IEMOCAP** compared to state-of-the-art SER models; a user study demonstrates improved immersion and satisfaction.

Strengths

Novelty with practical significance

- Clearly identifies limitations of existing subtitle systems (lack of emotional expression and personalization) and provides meaningful contributions both technically and practically.

Well-structured methodology

- Cross-attention design for multimodal fusion is clearly structured and well-motivated.

Major Weaknesses

Limited analysis of personalization impact

- While personalized adaptation is introduced, the actual emotional effects are insufficiently analyzed.
- For example, subjective differences in color-emotion associations across cultural groups are not systematically validated.

Insufficient technical depth

- Although the combination of Wav2Vec2 and BERT is described, the paper lacks ablation studies or interpretability analyses to clarify their specific contributions within the system.

Minor Weaknesses

- Personalization mainly focuses on static attributes (age, culture) without considering dynamic aspects such as real-time user emotion or interactive feedback.
- Comparative experiments emphasize SER model performance but lack direct comparisons with other emotion-enhanced subtitle systems.

Rating: 4: Ok but not good enough - rejection
Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Add: **Ethics Chair Author Comment**

AI Review

AI Review by Program Committee AI 27 Aug 2025, 14:30 (modified: 16 Sept 2025, 04:07)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee AI, Program Committee Submitted, Authors
Revisions (/revisions?id=UvUPJiUv6e)

Review:
Title: Sentimentogram: A Personalized Subtitle System Using Multi-modal Emotion Recognition and Visualization

Synopsis of the paper The paper presents Sentimentogram, a system that augments conventional subtitles with emotion-aware styling and user personalization. It fuses acoustic features from Wav2Vec2 and textual features from BERT via self- and cross-attention to classify emotions, maps these to a valence–arousal–dominance space, and uses the mapping to drive color and typographic choices. A personalization layer adapts rendering by age (e.g., slang vs. censorship), culture (culture-specific color–emotion associations), and playback speed (emoji cues and character-per-second control). The system is evaluated with an IEMOCAP-based benchmark for emotion recognition and with a user study comparing standard and Sentimentogram subtitles.

Summary of Review The paper addresses a timely and practically meaningful gap by operationalizing multimodal emotion recognition into an end-user-facing, adaptive subtitle interface. The interface-level integration of emotion inference, color/typographic rendering, and culturally aware personalization is original and aligns with accessibility goals. However, there are multiple technical inaccuracies (e.g., the emoji-weighting function; attention notation; label-set inconsistencies; VAD formulation vs. implementation) and important missing details (training protocol, alignment, baselines, and user study design) that limit reproducibility and the strength of claims. The evaluation omits several closely related state-of-the-art comparators under matched protocols, and the user study lacks essential methodological information and controlled analysis of each personalization axis. With corrections, fuller baselines, and a more rigorous evaluation, the contribution could be compelling for both affective computing and human–computer interaction.

Strengths

- Clear problem framing and system concept
  - The paper convincingly motivates the need for emotion-aware, context-sensitive subtitles beyond static speech-to-text, and articulates a pipeline that connects recognition, affect mapping, and personalization.
  - The emphasis on cultural adaptation (color–emotion associations) and playback-speed-aware presentation is a thoughtful extension of standard speech emotion recognition toward accessible, user-centered media experiences.
- Multimodal fusion architecture with ablation
  - The design—self-attention over concatenated audio and text features followed by bidirectional cross-attention—is a reasonable and increasingly common approach in multimodal affect recognition.
  - The ablation (text-only, audio-only, naive concatenation, proposed) shows a substantial gain from attention-based fusion, which supports the claim that cross-modal interactions matter.
- VAD-driven visualization and cultural grounding
  - Mapping predicted emotion to a continuous valence–arousal–dominance representation to parameterize color/typography is well-motivated by affective science, and the associations are cross-validated against large-scale color–emotion surveys.
  - The proposed cultural knowledge graph and rule-based adaptation reflect relevant work in culturally adaptive interfaces.
- Dual evaluation modalities
  - Combining benchmark SER results (IEMOCAP) with a user study on subtitle perception speaks to both technical performance and perceived utility.

Weaknesses

- Technical inaccuracies and inconsistencies in core methodology
  - Emoji-weighting function contradicts the intended behavior. The paper claims the emoji weight W_{emoji} should approach 1 at high playback speeds, but defines $W_{emoji}(s) = \max(0, \min(1, \frac{s_{min}-s}{s_{max}-s_{min}}))$, which decreases in s and clamps to 0 for $s \geq s_{min}$. As written, this cannot realize the described behavior.
  - Attention equations are nonstandard and under-specified. The manuscript writes $H' = \mathrm{SelfAttention}(X_q, X_v, X_k)$ without defining how Q, K, and V are formed or their projections. The residual/normalization is written as "Layer," but the intended operation (e.g., LayerNorm) is not specified. Dimensionality notation is inconsistent (N vs. N_s, N_t).
  - Emotion label sets are inconsistent across the pipeline. The figures and text reference: (a) six basic emotions (anger, disgust, fear, sad, surprise, joy), (b) a seven-class set including neutral in the VAD table, and (c) IEMOCAP six-class evaluation (happiness, sadness, neutral, anger, excitement, frustration). The mapping between these sets is not provided, creating ambiguity in training targets, VAD lookup, and rendering.
  - VAD formulation vs. implementation mismatch. The equation $E = (V, A, D) = f(X; W)$ suggests regression of continuous VAD from input X, yet the implementation appears to classify into discrete emotions and then look up VAD values from a static lexicon. There is no defined regression objective that would make f a learned mapping.
  - Typography mapping discrepancy. The abstract claims adjusting font size by "emotional intensity," but the methods map dominance to lightness and font weight; no font-size rule is defined. Additionally, the emoji-insertion example ("I'm so happy 😊 today!") conflicts with the stated policy to prepend the emoji to the beginning of the subtitle.
  - Minor math/notation slips. For example, the pooling/logit equation uses $\bar{h}$ earlier, but the subsequent line in one place refers to $z = hW_c + b_c$ (missing the bar), and $F_{\text{Audio}}/F_{\text{Text}}$ are first defined with length N and later with N_s/N_t.
- Missing implementation and training details that affect reproducibility
  - Sequence handling and alignment: how audio frames (N_s) and BERT tokens (N_t) are temporally aligned (or whether they are not), whether positional and modality/type embeddings are used, and how masking is handled for variable-length inputs is not described.
  - Backbones and hyperparameters: the exact Wav2Vec2 and BERT variants, tokenization, sampling rate and feature framing, optimizer, learning rates, batch size, training schedule, early stopping, number of runs and seeds, and compute resources are not reported.

- Domain adaptation with Condensed Movies is not quantified: there is no ablation or protocol indicating how the dataset is incorporated and what performance impact it yields.
- It is unclear whether the text branch uses ground-truth transcripts or Whisper outputs; if Whisper is used, the impact of automatic speech recognition noise on the text modality is not assessed.

- Evaluation gaps and unsubstantiated state-of-the-art positioning
  - Baselines in Table 2 are drawn from heterogeneous settings and exclude several closely related and strong contemporaries under matched protocols:
    - emotion2vec for the acoustic branch (Ma et al., 2024),
    - correlation-aware cross-attention fusion (Shi & Huang, 2023),
    - teacher-leading multimodal distillation (Yun et al., 2024),
    - continuous VAD prediction with Wav2Vec2+BERT (Zhang et al., 2024),
    - contrastive decomposition/fusion approaches (Peng et al., 2024; Yang et al., 2023).
  - Metrics are limited to weighted accuracy and weighted F1. In class-imbalanced emotion recognition, unweighted accuracy (or unweighted average recall) and macro-F1 are standard. Per-class precision/recall and a confusion matrix are missing.
  - IEMOCAP protocol is unspecified: speaker-independent folds, label mapping/merging (e.g., happy and excited), and whether frustration is included are unclear. Without a precise protocol, the comparability of the reported numbers is difficult to judge.
- User study methodology is insufficiently described
  - The study omits sample size, demographics, language proficiency, design (within vs. between subjects), counterbalancing, stimuli, instructions, apparatus, and statistical analyses (tests, confidence intervals, effect sizes).
  - The study does not isolate contributions of specific personalization axes (e.g., color-only, typography-only, age-based text transformation, emoji-speed cue, cultural mapping) and does not test cultural adaptation with culturally diverse cohorts or account for color-vision deficiencies and contrast accessibility.
- Practical and design considerations not fully addressed
  - Color/typography choices may reduce legibility over varied video backgrounds and for users with color-vision deficiencies; the paper does not report contrast ratios or accessibility compliance.
  - The age-based slang substitution and profanity filtering may affect comprehension and semantic fidelity but are not evaluated for potential adverse effects.
  - Real-time feasibility and latency are not reported, despite the application context.

Suggestions for Improvement

- Correct and clarify the technical formulation
  - Fix the emoji-weight function to be monotone increasing in playback speed, for example:

    $$W_{emoji}(s) = \max\left(0, \min\left(1, \frac{s - s_{min}}{s_{max} - s_{min}}\right)\right)$$

    with $s_{max} > s_{min}$, and ensure the example matches the stated prepending policy (e.g., "😊 I'm so happy today!").
  - Standardize the attention notation and define all projections. Specify how Q, K, V are obtained from X (e.g., linear projections), the number of heads, dimensions, positional and modality embeddings, masking, and residual/normalization details.
  - Unify the emotion label set across training, evaluation, and visualization. Explicitly define the IEMOCAP protocol (e.g., merging happy/excited, handling frustration) and provide a clear mapping from the evaluation labels to the rendering palette and VAD table.
  - Clarify the VAD objective. If VAD is a lookup from predicted classes, rewrite E as a deterministic mapping from discrete labels to a VAD triplet. Alternatively, if VAD regression is intended (as in Zhang et al., 2024), define the regression loss, targets, and training setup.
  - Fully specify typographic mappings: if "emotional intensity" implies font size, define the mapping (e.g., $A \mapsto$ size range), and keep $D \mapsto$ lightness/weight, with parameter ranges and clamping.
- Provide complete implementation and training details to enable replication
  - Describe sequence alignment between audio frames and text tokens (or justify late fusion without alignment), the use of positional/modality embeddings, and variable-length masking. Report backbones (e.g., Wav2Vec2-Base vs. Large; BERT-Base vs. Large), preprocessing, optimizer, learning-rate schedule, batch size, epochs, early stopping, seeds, and compute.
  - Define the IEMOCAP splits and protocol and report mean ± standard deviation over multiple runs. Add unweighted accuracy or unweighted average recall and macro-F1, per-class scores, and a confusion matrix.
  - Clarify whether Whisper transcripts or ground-truth transcripts are used; if Whisper is used, quantify the impact of ASR noise on the text branch and overall performance (e.g., by comparing to ground-truth transcripts), in line with findings on ASR sensitivity (Feng et al., 2024).
  - Quantify the contribution of the Condensed Movies fine-tuning with an ablation (with/without domain adaptation), and consider reporting cross-corpus generalization.
- Strengthen baselines and analysis under matched protocols
  - Add strong, closely related baselines:
    - Use emotion2vec features for the audio branch (Ma et al., 2024).
    - Compare to correlation-aware cross-attention fusion (Shi & Huang, 2023).
    - Evaluate teacher-leading distillation from text to audio (Yun et al., 2024) to bolster robustness in noisy media.
    - Include continuous VAD prediction with Wav2Vec2+BERT (Zhang et al., 2024) as a methodologically close comparator to the VAD-to-visualization pipeline.
    - Explore contrastive decomposition/aggregation (Peng et al., 2024; Yang et al., 2023) for more structured fusion objectives.
  - Provide an error analysis: common confusions (e.g., anger vs. frustration), modality dominance in successes/failures, and robustness to background noise, speaker variability, and ASR errors.
- Substantially expand the user study methodology
  - Report participants (number, demographics, language proficiency), inclusion criteria, stimuli, apparatus, design (within/between), counterbalancing, and statistical analyses (tests, confidence intervals, effect sizes). Pre-registering the analysis plan would further strengthen validity.
  - Isolate the impact of each personalization axis via controlled conditions (e.g., baseline vs. +color-only vs. +typography-only vs. +age-based text vs. +emoji-speed vs. +culture mapping). Specifically validate cultural adaptation with culturally diverse cohorts. Screen for color-vision deficiencies and report contrast compliance.
  - Add objective measures: reading speed, comprehension accuracy, and cognitive load, and analyze potential trade-offs (e.g., whether emojis distract at normal speeds).
- Broaden the technical scope and practical considerations

- Consider integrating visual facial cues and speaker conditioning when available, which have shown benefits in conversational settings (Zheng et al., 2023).
- Explore model-side robustness that complements interface-side personalization: topic/style conditioning (Qiu et al., 2023), sentiment-knowledge-enhanced self-supervised pretraining (Qian et al., 2023), or cross-modal distillation (Yun et al., 2024).
- Report runtime, latency, and memory usage for real-time rendering and the cost of personalization components. Discuss deployment considerations (e.g., batching, streaming inference, and fallbacks to a "safe palette" with sufficient contrast).
- For color/typography, adopt accessibility guidelines (e.g., WCAG contrast), add outlines or drop-shadows, and offer user controls to opt out of slang substitution and emoji cues.

- Position the contribution more precisely
  - Calibrate the "first comprehensive effort" claim by engaging with recent multimodal fusion and interface-level work. Clearly articulate architectural differences relative to cross-attention-based fusion (Shi & Huang, 2023) and continuous VAD prediction (Zhang et al., 2024), and explain why the chosen design offers practical advantages for subtitle rendering.

References

- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33.
- Bain, M., Nagrani, A., Brown, A., & Zisserman, A. (2020). Condensed Movies: Story-based retrieval with contextual embeddings. In Proceedings of the Asian Conference on Computer Vision.
- Busso, C., Bulut, M., Lee, C. M., Kazemzadeh, A., Mower, E., Kim, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42, 335–359.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.
- Feng, S., Sun, G., Lubis, N., Wu, W., Zhang, C., & Gasic, M. (2024). Affect recognition in conversations using large language models. In Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 259–273). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.sigdial-1.23 (https://doi.org/10.18653/v1/2024.sigdial-1.23)
- Gu, Y., et al. (2025). Scene-Speaker Emotion Aware Network: Dual network strategy for conversational emotion recognition. Electronics, 14(13), 2660.
- Jonauskaite, D., Abu-Akel, A., Dael, N., et al. (2020). Universal patterns in color–emotion associations are further shaped by linguistic and geographic proximity. Psychological Science, 31(10), 1245–1260.
- Li, Y., et al. (2024). Tracing intricate cues in dialogue: Joint graph structure and sentiment dynamics for multimodal emotion recognition. arXiv:2401.01495.
- Ma, Y., et al. (2023). A transformer-based model with self-distillation for multimodal emotion recognition in conversations. IEEE Transactions on Multimedia.
- Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., & Chen, X. (2024). emotion2vec: Self-supervised pre-training for speech emotion representation. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 15747–15760). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.931 (https://doi.org/10.18653/v1/2024.findings-acl.931)
- Peng, C., Chen, K., Shou, L., & Chen, G. (2024). CARAT: Contrastive feature reconstruction and aggregation for multi-modal multi-label emotion recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 38(13), 14581–14589. https://doi.org/10.1609/aaai.v38i13.29374 (https://doi.org/10.1609/aaai.v38i13.29374)
- Qian, F., Han, J., He, Y., Zheng, T., & Zheng, G. (2023). Sentiment knowledge enhanced self-supervised learning for multimodal sentiment analysis. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 12966–12978). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.821 (https://doi.org/10.18653/v1/2023.findings-acl.821)
- Qiu, S., Sekhar, N., & Singhal, P. (2023). Topic and style-aware transformer for multimodal emotion recognition. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 2074–2082). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.130 (https://doi.org/10.18653/v1/2023.findings-acl.130)
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning.
- Shi, T., & Huang, S.-L. (2023). MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (pp. 14752–14766). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.824 (https://doi.org/10.18653/v1/2023.acl-long.824)
- Wu, Z., et al. (2024). Beyond silent letters: Amplifying LLMs in emotion recognition with vocal nuances. arXiv:2407.21315.
- Yang, J., Yu, Y., Niu, D., Guo, W., & Xu, Y. (2023). ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (pp. 7617–7630). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.421 (https://doi.org/10.18653/v1/2023.acl-long.421)
- Yin, B., et al. (2025). Adaptive graph learning with multimodal fusion for emotion recognition in conversation. Biomimetics, 10(7), 414.
- Yun, T., Lim, H., Lee, J., & Song, M. (2024). TelME: Teacher-leading multimodal fusion network for emotion recognition in conversation. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 82–95). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.5 (https://doi.org/10.18653/v1/2024.naacl-long.5)
- Zhang, E., Trujillo, R., & Poellabauer, C. (2024). The MERSA dataset and a transformer-based approach for speech emotion recognition. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (pp. 13960–13970). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.752 (https://doi.org/10.18653/v1/2024.acl-long.752)
- Zheng, W., Yu, J., Xia, R., & Wang, S. (2023). A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (pp. 15445–15459). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.861 (https://doi.org/10.18653/v1/2023.acl-long.861)
- Valdez, P., & Mehrabian, A. (1994). Effects of color on emotions. Journal of Experimental Psychology: General, 123(4), 394.